ADOC : Agrégation des DOnnées par Comparaisons stochastiques

Mots clés : Agrégation, Comparaison stochastique, Monotonie stochastique, Algorithmique, Complexité.

Résumé scientifique du projet : Le projet a pour but de fournir des algorithmes d'agrégation de données de grande taille qui permettent d'apporter des garanties, au sens d'un ordre stochastique, après traitement. Ces méthodes ont été déjà définies pour l'ordre stochastique fort afin d'étudier des traces de trafic de réseau IP. Nous nous proposons d'étendre ces techniques à des ordres plus complexes pour permettre d'analyser des données de grandes tailles, en construisant des agrégats qui garantissent des bornes. Le projet regroupe des informaticiens et des probabilistes appliqués, mais il comprend aussi des sociologues et des spécialistes de l'énergie.

Les dispositifs de collecte permettent de construire des masses de données que nous ne pouvons exploiter par des méthodes conventionnelles. De nombreux traitements ne peuvent être réalisés de façon efficace sur des masses de données trop larges à cause de leur complexité. Plutôt que sélectionner une partie de ces données (pas forcément significative), nous nous proposons de les agréger par des algorithmes qui fournissent des bornes stochastiques (au sens d'ordres stochastiques que nous définirons plus loin). Le point fort et original de cette approche est que cette compression de données peut être effectuée en garantissant des bornes sur un traitement dont on prouvera la monotonie par rapport à l'ordre considéré. L'agrégation proposée est donc dépendante du traitement postérieur.

Le cadre de la méthode est l'étude et l'utilisation de données constituant des échantillons indépendants issus d'une distribution stationnaire pendant la période de mesure. On suppose que ces échantillons sont trop nombreux pour s'en servir directement et qu'un prétraitement est nécessaire pour que les algorithmes ultérieurs travaillent de façon efficace et robuste sur des données agrégées.

Un ordre stochastique est un ordre sur les distributions de probabilités. Une façon de les définir (ce n'est pas la seule) consiste à considérer un cône de fonctions F et à définir qu'une distribution X est plus petite qu'une distribution Y, si l'espérance de F(X) est plus petite que l'espérance de F(Y), pourvu que ces espérances existent pour toutes les fonctions du cône. Ainsi l'ordre stochastique fort est associé aux fonctions croissantes, et l'ordre convexes aux fonctions convexes. Nous considérons des échantillons et ceci constitue une distribution discrète. La complexité de cette distribution est de façon naturelle, la taille du support (le nombre de valeurs distinctes).

L'idée principale est double :

- Utiliser des ordres stochastiques pour agréger les données dans une distribution avec un plus petit support mais qui constitue une borne au sens de l'ordre stochastique considérée. Pour les ordres stochastiques classiques, il faut développer une approche algorithmique permettant d'agréger les données tout en garantissant l'ordre. Ces algorithmes doivent être suffisamment rapides pour être appliquées à des données de grande taille.

- Étudier la monotonie des traitements au sens de l'ordre retenue pour l'agrégation. En effet, si les traitements sont monotones pour cet ordre, alors la comparaison entre la distribution d'entrée et la distribution agrégée sera également vraie pour les résultats de sortie. Le traitement algorithmique des données agrégées fournira donc une garantie au sens de l'ordre stochastique pour les résultats.

Il ne s'agit pas d'une heuristique de sélection ou de filtrage mais de constructions algorithmiques des valeurs agrégées qui apportent une garantie sur le résultat final, grâce à la compatibilité entre l'ordre utilisé pour l'agrégation et l'ordre employé pour caractériser la monotonie.

Donnons immédiatement un exemple issu de nos travaux sur l'ordre stochastique fort. Supposons que l'on agrège une distribution d'entrée composée de N échantillons par leur K agrégats tout en faisant en sorte que la distribution sur les K éléments soit une borne stochastique de la distribution empirique fournie par les N échantillons. Un point de vue intuitif de l'ordre stochastique fort (<st) est que les fonctions de répartition FX et FY ne se croisent pas :

X <st Y si et seulement si FY ≤ FX

Supposons, par exemple, que ces données représentent des délais aléatoires dans un réseau de transport. Si on calcule à partir de ces données agrégées une fonction combinant des opérateurs croissant convexes (comme le délai minimal de bout en bout dans le réseau, variable aléatoire qui s'exprime à partir des opérateurs Min et Somme, tous deux convexes et croissants, et des distributions empiriques) alors on obtiendra à partir des données agrégées une borne inférieure au sens de l'ordre stochastique fort de la distribution réelle de la variable aléatoire. On a donc issu une monotonie à la fois pour les ordres stochastiques forts (à cause des fonctions croissantes), et convexes (à cause des fonctions convexes). Agréger les échantillons permet donc d'analyser la distribution de la variable aléatoire, de façon plus rapide, en fournissant une borne du résultat exact (qui n'est peut-être pas calculable en un temps raisonnable).

Dans le cadre de l'ordre stochastique fort, on a pu démontrer qu'il était suffisant de chercher un sous ensemble de cardinalité K parmi les N atomes de la distribution empirique. On a donc un problème de nature combinatoire pour chercher le meilleur sous-ensemble de K atomes parmi N. Dans [3] nous avons fourni plusieurs solutions à ce problème : plusieurs algorithmes gloutons et une solution qui minimise la différence de deux espérances de récompenses sur les deux distributions. Ce dernier algorithme repose sur la programmation dynamique.

Nous avons développé cette méthode, au cours de la thèse de F. Ait Salaht, afin d'analyser des réseaux IP, pour lesquels existent de volumineuses mesures d'arrivées. Par exemple, la trace MAWI [D1] contient les mesures de trafic IP sur un lien trans-Pacifique à 150Mbps obtenues pendant 1 heure en 2007. La trace comprend un nombre trop important d'échantillons pour décrire de façon efficace le processus d'arrivée du trafic que l'on utilise dans une analyse stationnaire d'un élément de réseau (la complexité numérique est de l'ordre du cube de la taille, dans ce cas précis). On a donc remplacé la distribution par une version agrégée au sens de l'ordre stochastique fort et montré la monotonie, au sens de cet ordre, des modèles de plusieurs éléments de réseaux. Les résultats les plus significatifs sont repris dans [1] et [2]. On obtient par cette méthode des garanties de performance au sens de l'ordre stochastique fort.
D'un point de vue théorique, les méthodes d'agrégation généralisent le principe de fusion de distribution de Elton et Hill.

Cette méthode a également été appliquée dans le cadre de l'ANR MARMOTE (ANR-12-MONU-00019) pour comparer au sens de l'ordre stochastique fort, les données de temps de vie résiduel [4]. Un logiciel d'analyse des Arbres de Fautes Dynamiques s'appuyant sur cette approche est en cours de réalisation au LACL. Il est à noter que cette ANR a fait suite à un projet PEPS CNRS sur l'approche temps parallèle et la monotonie en simulation.

Les ordres stochastiques ont été étudiés dans de nombreux domaines d'applications des probabilités : fiabilité, assurance, finance, performance des réseaux et des systèmes de production. Par contre, leur emploi dans le cadre du big data nous semble trop limité pour le moment par manque d'une approche algorithmique véritablement efficace et aussi, dans certains cas, par manque d'une méthodologie de traitements des données. L'absence de cette approche algorithmique constitue le verrou principal que nous souhaitons lever dans ce projet.

Le projet est articulé autour des deux axes et guidé par des domaines applicatifs puisque l'ordre stochastique à employer dépend de la monotonie de l'application terminale qui va donc dépendre du domaine. Dans tous les cas, il s'agit d'associer via un ordre stochastique, des algorithmes d'agrégation qui vérifient la relation d'ordre sur les données produites et des algorithmes de traitement qui doivent être monotones au sens de l'ordre. Ces algorithmes de traitement existent déjà mais leurs propriétés relatives à la monotonie sont en général inconnues alors que les algorithmes d'agrégation associés à un ordre stochastique sont à concevoir.

Agréger les données est aussi une réponse simple au besoin d'anonymat et de « privacy ». Les échantillons qui proviennent d'une agrégation regroupent plusieurs individus et ne permettent pas d'affecter à un seul d'entre eux, les caractéristiques de l'agrégat. De plus, on connaît déjà des algorithmes distribués permettant de calculer la moyenne des échantillons sans que les participants connaissent les données des autres participants. La moyenne est une des stratégies d'agrégation qui interviendra pour l'ordre convexe (voir ci dessous). Généraliser ces algorithmes au calcul d'autres moments (au sens des probabilités) est simple. Il reste à travailler sur des algorithmes de calcul permettant de modifier de façon anonyme des distributions. Calculer des agrégats avec de tels algorithmes permet d'avoir des

données anonymisées où les échantillons correspondent à des individus virtuels regroupant une partie de la population.

Plus précisément, on traitera des questions suivantes :

1) développer les algorithmes d'agrégation pour les ordres « convexe » et « convexe croissant ». L'ordre convexe permet de considérer une distribution qui aura la même moyenne mais dont la variance sera réduite (pour une borne inférieure) ou augmentée (pour une borne supérieure). Ces ordres permettent de borner l'aversion au risque (à moyenne constante) car la variance des échantillons traduit souvent un risque (au sens large) pour le phénomène étudié. Nous comptons employer ces ordres (associés à l'ordre stochastique fort) pour analyser des phénomènes d'attente représentés par des files d'attente ou des réseaux de Petri. Dans ces modèles, il est classique de faire apparaître des séquences récursives stochastiques (par exemple l'équation de Loynes) et il suffit de démontrer les propriétés de croissance ou de convexité des opérateurs impliqués dans ces séquences pour démontrer la monotonie au sens de ces ordres stochastiques. Les exemples d'application sont des analyses des réseaux de télécom dans la poursuite de nos travaux précédents à partir des traces MAWI et CAIDA [D1, D5] mais aussi l'analyse des phénomènes d'attente des patients dans des hôpitaux [D3] (en particulier les urgences) pour proposer des stratégies robustes de planification pour les agents. Nous avons bon espoir d'obtenir des algorithmes linéaires dans la taille pour construire des bornes inférieures ou supérieures au sens de l'ordre convexe. Pour agréger une distribution au sens de l'ordre convexe, il est assez naturel de calculer des agrégats qui sont des moyennes des individus. Nous avons prouvé que la distribution agrégée est une borne inférieure au sens de l'ordre convexe (également appelé ordre « stop-loss order » en modélisation financière). Pour calculer une borne supérieure pour le même ordre, nous devons diminuer la taille de l'échantillon tout en augmentant sa variance. Diverses solutions ont déjà été proposées et il faut chercher la plus efficace d'entre elles, en terme de complexité mais aussi pour la précision de la borne.
2) Créer une description algorithmique pour les distributions discrètes pour les ordres stochastiques « likelihood ordering » et « submodular ordering ». Le premier de ces ordres est utilisé en fiabilité et pour l'analyse des risques et de la disponibilité. La deuxième méthode de comparaison correspond aux propriétés de sous modularité qui sont souvent employées en optimisation et en tarification. Nous comptons employer ces deux méthodes pour étudier les phénomènes de variation de production pour les systèmes de production solaire [D2] et leur impact tant pour la stabilité et le contrôle de la grille que pour la tarification [6,7].
3) Nous allons développer des modèles très simples (en partant des données [D2]) pour un agrégat des utilisateurs. Pour cela, nous allons utiliser les ordres st et l'ordre convexe afin de borner le risque lié à un pic de consommation électrique. Les mêmes données seront utilisées pour développer des modèles des charges pour pouvoir faire du pilotage automatique de la consommation électrique des utilisateurs afin de minimiser le coût de l'électricité, dans le contexte des programmes de l'effacement de la demande dans les périodes du pic de consommation, ou encore pour contribuer à des services système (cf. [6], [7]).

4) Étudier les ordres de dépendance pour les variables aléatoires bivariées et les techniques algorithmiques pour construire des agrégations d'échantillons. En sociologie, les études quantitatives reposent souvent sur l'analyse statistique de données dont la première étape consiste à calculer une matrice de similarité ou de distance. Les algorithmes (composantes principales, clustering, décomposition de rang faible) qui en découlent, ont en général une complexité quadratique ou cubique, et il est difficile d'analyser de très larges échantillons. Les variables sont bien sûr dépendantes et les méthodes d'ordre stochastique doivent prendre en compte certaines propriétés de dépendance. Dans la pratique, on distingue les ordres associés à la dépendance positive (par exemple lié à une corrélation positive, telle que la notion de variables associées) et les ordres associés à la dépendance négative. On se propose de donner une version algorithmique de certains de ces ordres et d'étudier dans quelle mesure ils permettent de remplacer de grandes matrices de similarité par des matrices de plus petite taille sur des individus agrégés tout en garantissant de conserver des résultats qualitatifs.

5) Cette approche de dépendance positive a aussi des applications en fiabilité où on dispose de mesures sur les distributions de durée de vie des éléments de base. L'approche classique consiste à les composer (par exemple dans un arbre de fautes, statique ou dynamique) en supposant l'indépendance. Dans la réalité, les pannes et les fautes semblent souvent positivement corrélées car elles sont liées à des événements déclencheurs communs ou dans une cascade. Il est donc nécessaire de savoir composer les données de vies empiriques provenant des mesures sous des hypothèses de dépendance positive et de savoir simplifier ces distributions empiriques pour que leur taille soit compatible avec des traitements numériques [4,5].

Référence :

[1] « Performance Analysis of a Queue by Combining Stochastic Bounds, Real Traffic Traces and Histograms, » The Computer Journal, 2016, F. Aït Salaht, H. Castel Taleb, J.M. Fourneau, et N. Pekergin.

[2] « A Bounding Histogram Approach for Network Performance Analysis », F. Aït-Salaht, H. Castel-Taleb, J.-M. Fourneau, et N. Pekergin, 10th IEEE International Conference on High Performance Computing and Communications, HPCC, China, pp 458--465, 2013, Best Paper Award.

[3] « Accuracy vs. Complexity: the stochastic bound approach », 11th International Workshop on Discrete Event Systems (WODES2012) N8, pp 178-192, 2012, F. Aït Salaht, J. Cohen, H. Castel Taleb, J.M. Fourneau, et N. Pekergin.

[4] « A Numerical Analysis of Dynamical Fault Trees based on Stochastic Bounds », IEEE QEST, 2015, J.M. Fourneau, et N. Pekergin.

[5] « Computing stochastic bounds of network distributions of time before failure », IEEE RNDM 2015, J.M. Fourneau

[6] « Ancillary Service to the Grid Using Intelligent Deferrable Loads », S. Meyn, P. Barooah, A. Busic, Y. Chen, J. Ehren, IEEE Trans. Automat. Control. 60 (11): pp 2847 - 2862. 2015, également dans PGMO DAYS 2015.

[7] « Distributed Randomized Control for Demand Dispatch», A. Busic, S. Meyn. Mar 2016. 55th IEEE Conference on Decision and Control, Dec 2016.

Données étudiées ou en cours d'étude

[D1] MAWI trace : « traffic data repository at the wide project », Sony, KC, Cho K, 2000, Usenix Annual Technical Conference.

[D2] PECAN Street, données provenant de plus de 250 maisons dans les environs d'Austin (Texas) décrivant la consommation et la production électrique (panneaux solaires). Les données sont échantillonnées toutes les minutes et proviennent des panneaux mais aussi d'une vingtaine d'autres circuits. https://dataport.pecanstreet.org

[D3] Hôpital : Données d'attente sur les service d'urgence et plusieurs services hospitaliers, RambamData sur le site : http://seeserver.iem.technion.ac.il/see-terminal.

[D4] Enquêtes nationales Transports . Enquêtes périodiques réalisées par l'INSEE auprès des ménages quant à leurs modes et habitudes de déplacements, les enquêtes nationales transports se distinguent par un protocole particulièrement complexe (carnet de déplacements, carnet de véhicules, questionnaires individuels et ménages). Sous-utilisées par les chercheurs en sciences humaines, elles sont tout à fait susceptibles d'être valorisées dans le cadre du projet proposé, notamment dans la mise en place de classifications automatiques, particulièrement utiles pour les sociologues.

[D5] CAIDA : Traces of OC48 links at Ames Internet eXchange (aix) (april 24, 2008), internet data measurement catalog, http://imdc.datacat.org.